Exploring self-supervised learning techniques for hand pose estimation

Aneesh Dahiya
Department of Computer Science
ETH Zurich, Switzerland
adahiya@student.ethz.ch

Adrian Spurr
Department of Computer Science
ETH Zurich, Switzerland
adrian.spurr@inf.ethz.ch

Otmar Hilliges
Department of Computer Science
ETH Zurich, Switzerland
otmar.hilliges@inf.ethz.ch

Abstract

3D hand pose estimation from monocular RGB is a challenging problem due to significantly varying environmental conditions such as lighting or variation in subject appearances. One way to improve performance across board is to introduce more data. However, acquiring 3D annotated data for hands is a laborious task, as it involves heavy multi-camera set up leading to lab-like training data which does not generalize well. Alternatively, one could make use of unsupervised pre-training in order to significantly increase the training data size one can train on. More recently, contrastive learning has shown promising results on tasks such as image classification. Yet, no study has been made on how it affects structured regression problems such as hand pose estimation. We hypothesize that the contrastive objective does not generalize easily to such downstream task due to its inherent invariance property stemming from the and instead propose a relation objective, promoting equivariance. Our goal is to perform extensive experiments to validate our hypothesis.

1 Introduction

Given a monocular RGB image, estimating the location of hand joints is a challenging structured regression problem. Amongst others, conditions that significantly contribute to the difficulty are large diversity in backgrounds, lighting conditions and hand appearances, as well as self-occlusion.

One straightforward way of improving the performance of a learning-based model is to include more training data. However, acquiring 3D labeled data is laborious and expensive as it requires large lab-like setting whose data does not translate well to in-the-wild imagery [1, 2]. The community has been relying increasingly more on supplementary 2D annotated data to tackle this and demonstrated that inclusion of this additional data leads to better prediction accuracy. For example, [3] showed that one can outperform many supervised approaches by using weakly-supervised data more effectively via appropriate priors. Although easier to acquire, 2D annotations do not come for free. To tackle this, works exist [2] that use automatically generated 2D annotations with the help of OpenPose [4]. However, there is no guarantee that these poses are indeed correct and the accuracy one can achieve with such an approach is bounded by the performance of the OpenPose model.

Alternatively, one could resort to using unlabeled data directly with the help of self-supervision. Recently, approaches such as [5, 6] have shown that they are close to reaching parity or even

Pre-registration workshop NeurIPS (2020), Vancouver, Canada.

outperform supervised baseline models with the help of contrastive learning on tasks such as image classification. This raises an interesting question: *Does the contrastive self-supervised learning capability extend to structured regression tasks as well?* We hypothesize that features learned during contrastive-based training may not readily transfer to regression-based tasks, as the former results in features being *invariant* to the respective transformations. However, structured regression-based task require *equivariant* features. For example, given two images of the same hand, one being the rotated form of the other, the keypoints predicted on one hand should be the rotated version of the other. Yet, the objective function of contrastive learning encourages the features of both images to lie as closely as possible from one another, possibly inhibiting performance.

To tackle this, we propose a relative loss where the relative transformation from one image to the other is predicted. Our assumption is that this novel task pushes the model to learn a representation that is equivariant to the transformations applied. Coming back to our previous example of the two rotated hand poses, the relative loss requires the model to be able to predict the relative rotation between both the images. We hypothesize that doing so results in equivariant features, as the representation learned needs to be informative to infer the the applied transformation.

In this paper, we propose to explore self-supervised learning approaches for hand pose estimation by analyzing the currently prevalent method of contrastive learning. Our goal is to validate the hypothesis that the contrastive objective is not an effective way to leverage self-supervision and that by forcing the model to learn equivariant features, we can improve the performance of hand pose estimation approaches across the board. We want to compare our proposed loss with the original contrastive learning objective on the downstream task of hand pose estimation

We envision that the knowledge gained through the thorough evaluation of self-supervised methods in the context of structured regression problems will be valuable for communities such as hand and body pose. In the interest of reproducibility and contributing to the research community, we will be releasing the code and trained network model.

2 Related work

Self-supervised learning has gained interest in recent years as a form of unsupervised pre-training. Generally these rely on solving a pretext task which is not of interest to the actual task at hand. However, by solving the task, a good representation is learned as a by-product which can be used in downstream tasks.

Such pretext tasks can take any form. The most recent include Contrastive Multiview Coding (CMC) [7], Contrastive Predictive Coding (CPC) [8], simCLR [5], MoCo [6], whereas earlier works include [9, 10, 11, 12, 13, 14, 15, 16]. Alternatively, adversarial losses [17] can also be utilized for unsupervised representation learning [18, 19].

Authors in [20] tackle self-supervision by learning geometrically stable pixel level descriptors across a range of objects with probabilistic objective. Whereas in [21] and [22], authors estimate geometric features by predicting parameters for relative geometric transformation applied to the image and [23] estimates it by predicting one out of four angels used to rotate the input image. Differently, we propose to use geometric as well as appearance transformations. Lastly, none of the mentioned related work compares the contrastive with the pairwise relative loss formulation and does not report results on tasks such as hand pose.

In this paper, we focus on contrastive learning for self-supervision. However to the best of our knowledge, contrastive learning has not yet been applied to downstream tasks such as structured regression problems like that of hand pose estimation. One reasons for this could be that the resulting features may be invariant to the respective transformations, instead of equivariant. Our goal is to validate this hypothesis.

3 Methodology

Here we briefly recap the original contrastive formulation as was proposed by [5] and our proposed relative objective.

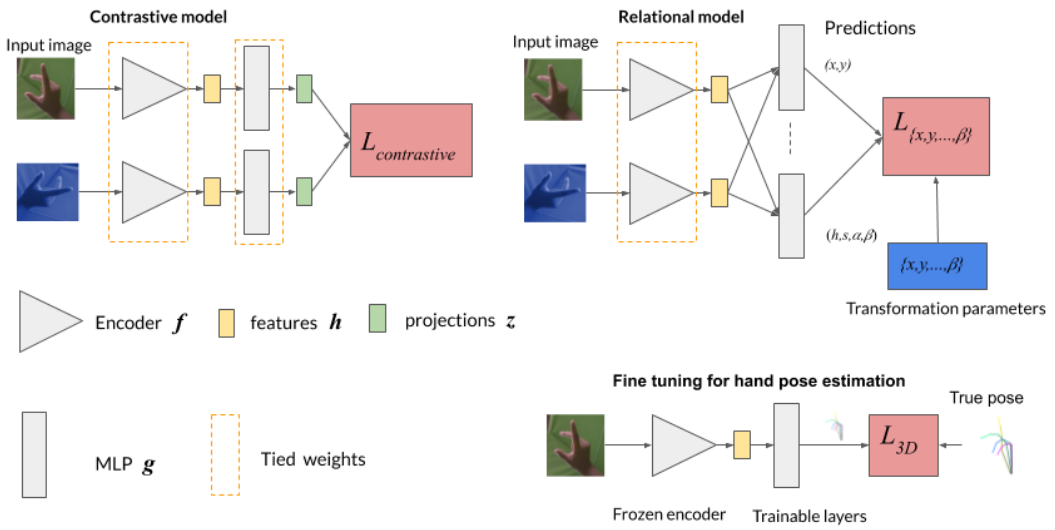


Figure 1: The pretraining phase of contrastive (left) and relative (right) models with fine tuning for hand pose estimation (bottom right). In the contrastive model, features generated by the encoder are passed through an MLP projection head to generate the projections on which the contrastive loss (Eq. 1) is computed. In the relative framework, the features generated by the encoder for a pair of transformed images are concatenated and passed through separate MLP heads to regress the relative transformation parameters. The relative loss is then computed on these predicted relative parameters using Eq. 3. After the pretext training phase, the encoders are frozen and the features generated from these encoders are then used to finetune a linear layer. The performance of the linear layer allows us to quantify the representation power of the features learned.

3.1 Recap on contrastive learning

We show an overview of the contrastive framework in Fig. 1 left. Contrastive learning enables a neural network f to learn features in an unsupervised manner by encouraging similar looking images to lie close in feature space. As such, it creates similar looking pairs of images by applying transformations $t_i : \mathcal{R}^n \rightarrow \mathcal{R}^n$ on a source image $\mathbf{x} \in \mathcal{R}^n$ and optimizing a neural networks weights to output similar features $f(t_i(\mathbf{x})) = \mathbf{h}_i$. These features are projected into a latent space $g(\mathbf{h}_i) = \mathbf{z}_i$ via a projection head g, on which the contrastive loss is applied to. To use the trained network f on a downstream task, the projection head is discarded and a linear classifier is trained on the features $\mathbf{h}$.

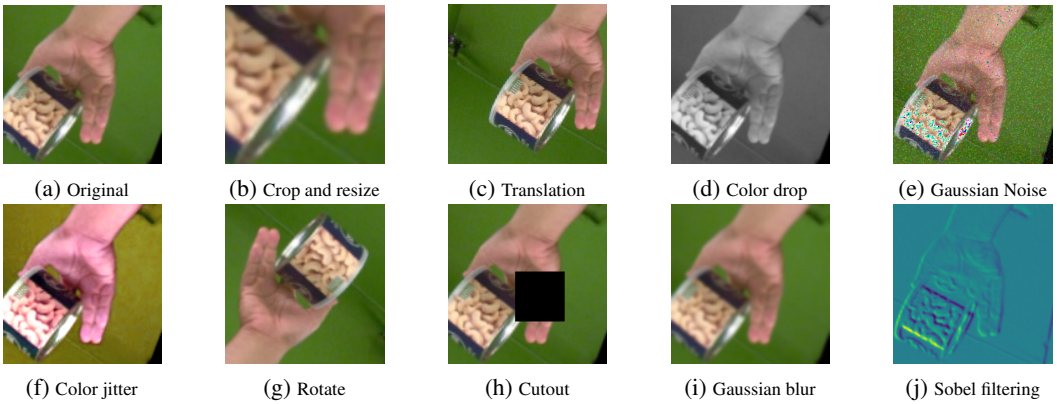
$$\sum_{i,j,i\neq j} -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{i\neq k} \exp \text{sim}((\mathbf{z}_i, \mathbf{z}_k)/\tau)} \tag{1}$$

Where $\text{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^T\mathbf{v}/||\mathbf{u}||_2||\mathbf{v}||_2$ computes a similarity and τ is a temperature parameter. Although impressive performance was achieved via this method, it is unclear how well these translate to structured regression problems. Although [5] report that they were capable of predicting the rotation angle with 67.6% accuracy, there are still issues: 1) It is unclear how the contrastive representation affects the structured regression-based downstream tasks 2) The classification was done by predicting an angle out of four. We hypothesise that there is more potential performance to be gained by reformulating the contrastive task to a relative one.

3.2 Proposal

Instead of contrasting an image pair, we propose to predict their relative transformations. Concretely, given a family of parameterized transformations $\mathcal{T}$ (e.g rotations), two randomly sampled parameters θ_i, θ_j (e.g rotation angles), we first compute the transformed sample pair via $\mathbf{x}_i = t(\mathbf{x}; \theta_i)$, where $t \in \mathcal{T}$, before passing it into the network f to obtain their respective features $\mathbf{h}_i$. These are fed into a transformation-specific projection head g_t to predict their relative transformation $\theta_{ij} = \theta_j - \theta_i$. As such, the objective changes from contrastive to relative. Hence, we reformulate Eq. 1 to following for one augmentation:

$$L_t = \sum_{i,j,i\neq j} ||\theta_{ij} - g_t(\mathbf{h}_i, \mathbf{h}_j)|| \tag{2}$$



(a) Original (b) Crop and resize (c) Translation (d) Color drop (e) Gaussian Noise

(f) Color jitter (g) Rotate (h) Cutout (i) Gaussian blur (j) Sobel filtering

Figure 2: Example of transformations used in this paper. Samples taken from FreiHAND [1].

In presence of $|\mathcal{T}|$ number of augmentations we minimize the loss described in Eq. 3, where the loss from each augmentation $t \in \mathcal{T}$ is scaled by a trainable parameter σ_t [24].

$$L = \sum_{k \in \mathcal{T}} (L_k / \sigma_k + \log \sigma_k) \tag{3}$$

As each task family $t \in \mathcal{T}$ is different, each g_t will be an independent network, but share the features $\mathbf{h}$ produced by the network f. Our hypothesis is that by predicting relative transformation parameters, the features learned will be equivariant to these transformations. This can be helpful for structured regression task where transformation of input also transforms the keypoints.

This intuition stems from the following. Given $\mathbf{x}$, we produce two samples $\mathbf{x}_1$ and $\mathbf{x}_2$ via $\mathbf{x}_i = t(\mathbf{x}; \theta_i)$. As such, our proposed objective is $||\theta_{21} - g_t(\mathbf{h}_2, \mathbf{h}_1)||_2$, where $\theta_{21} = \theta_2 - \theta_1$ is the target label. Transforming a new sample $\mathbf{x}_3$ via $\theta_3 = \theta_2 + \Delta\theta$ results in the new target label:

$$\theta_{31} = \theta_3 - \theta_1 = \theta_2 + \Delta\theta - \theta_1 = \theta_{21} + \Delta\theta. \tag{4}$$

Hence the target changes in accordance to the change in parameters. We postulate that this induces equivariance in the features h.

4 Experimental protocol

In order to provide fair comparison of our proposal with the contrastive loss, we will closely follow the experimental protocol outlined in [5]. The goal of the experiment section is to first verify which transformation benefits from which self-supervised loss. Next, we want to identify which composition results in the most beneficial feature representation. Lastly, we explore cross-dataset generalization and compare with fully supervised methods.

4.1 Protocol

We briefly describe the dataset and transformations used, metrics reported and the setting assumed for all experiments.

Datasets. We will benchmark our performance on two hand pose dataset. The first is FreiHAND (FH) [1] which contains 32560 samples of single hand pose with green screen backgrounds. Using synthetic background imagery, these are extended to $130k$ samples. The second is the InterHands2.6M (IH) dataset, of which we focus on the single-hand split which contains $688k$ samples.

Transformations. Following [5], we investigate the following transformations: crop, cutout, color jitter, sobel, noise, blur and rotate. A sample of these can be seen in figure 2. Due to chirality in hands, crop and flip is not used as an augmentation, instead we include random 2D translation of the hand as seen in figure 2c. All the images are pre-processed by cropping the hand and resizing it to 128×128 RGB image. Images in Fig. 2a are cropped from a larger image to isolate the hand, similar to [5].

In our proposed relative objective we investigate translation, color jitter and rotation as they have meaningful parameters that can be regressed. The translation parameters are (x, y) coordinates of crop box center. Rotation is characterized by an angle θ, around which the image is rotated. Color jitter is characterized by $h, s, (\alpha, \beta)$ parameters which change hue, saturation and value of the image pixels respectively. Augmentations like cutout and sobel filter are not included in the prediction since their parameters are trivial to regress. Relative augmentation parameters of gaussian blur, noise and cutout are ambiguous to predict, therefore are not included neither for our propose relative objective. We emphasize here that the images are still augmented with these augmentation, but we do not estimate their relative parameters.

Metrics. We report the mean per joint error (MPJE), as well as median on the downstream task of hand pose estimation. More specifically, given the self-supervised pre-trained network f, we follow the linear evaluation protocol [10, 8, 25, 26], where we train a linear layer on top of the frozen pre-trained network to regress the 2.5D hand pose representation. This allows us the quantify the feature representation learned in our proposed pretext task.

Setting. Following [27], we use a ResNet-18 [28] backbone network to facilitate training with bigger batch sizes as it was reported to improve performance [5, 6]. We use a 2-layer MLP projection head and a 128-dimensional latent space. All models are trained using the ADAM optimizer. Inspired by [5], the learning rate is scheduled using LARS [29] with an initial warmup phase to stabilize training for large batches. The learning rate is scaled using square root of the batch size n_{bs}, i.e. $lr = 0.0001 \times \sqrt{n_{bs}}$. For the downstream task of hand pose estimation, we discard the projection head and replace it with a linear layer. The optimal parameters are chosen using random grid search. For Sec. 4.5, we change the backbone network to that of [30], but keep the training scheme the same.

4.2 Data augmentation specific objective function

Before we attempt to investigate the ideal series of transformations, we need to answer a question: *Given our downstream task of regression, will all transformation yield a boost with our proposed relative objective? Could certain tasks pertain to the contrastive loss as opposed to the regressive loss?* For example, given the color augmentation where the color channels are augmented, it would perhaps be more beneficial to require the feature representation to be *invariant* as opposed to *equivariant*. In order to answer this question, we first perform an initial ablative study to inspect which transformation benefits from a relative objective as opposed to a contrastive one. To this end, we report the downstream task performance for each individual transformation, using either the contrastive or relative loss.

4.3 Data augmentation compositions

As was highlighted in [5], the composition of transformation operations is crucial to the final performance achieved in the downstream task. Since the downstream task here is regression, it is not clear if the same combination of transformations reported to be superior in [5] for classification will still remain as such in our downstream task. Therefore it is vital to determine which combination of transformations perform the best. To this end, we perform an exhaustive search. For each augmentation, we first pick the best performing pretext objective function, as determined in Sec. 4.2. Then, we inspect all possible combinations of augmentations and determine which composition performs best, based on the downstream task.

4.4 Cross-dataset generalization

Generalization is an important concept in any deep learning network. One simple way to cross the domain gap is to train on data of the target domain. However, often fully labeled data is only available in constrained lab environments. In this section, we want to explore our effectively self-supervised learning can be used to cross the domain gap. To this end, we perform self-supervised pre-training on IH and FH, but fine-tune the last linear layer only on FH. The final evaluation is done on IH to quantify if a reasonable improvement can be gained. In order to have a comparison, we train a fully supervised model solely on FH and compare the two results.

4.5 Comparison with supervised model

Given the best performing self-supervised objective and augmentation composition, we compare the performance against the current state-of-the-art hand pose estimator [30]. For this we replace the ResNet-18 encoder used in previous experiments with the hourglass model used in [30]. During the self-supervision phase, the 2D backbone network output is vectorized and passed through a non-linear projection layer, like in the prior experiments. During the downstream task training, we train the frozen network like that of [30]. The goal of this section is investigate how state-of-the-art models perform in the context of self-supervision.

References

[1] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images, 2019.

[2] Dominik Kulon, Riza Alp Güler, Iasonas Kokkinos, Michael Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild, 2020.

[3] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints, 2020.

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields, 2019.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020.

[7] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[8] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[9] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[10] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[11] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[12] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017.

[13] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[14] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

[15] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2015.

[16] Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2701–2710, 2017.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[18] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

[19] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In *Advances in Neural Information Processing Systems*, pages 10542–10552, 2019.

[20] David Novotny, Samuel Albanie, Diane Larlus, and Andrea Vedaldi. Self-supervised learning of geometrically stable features through probabilistic introspection, 2018.

[21] Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data, 2019.

[22] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. Convolutional neural network architecture for geometric matching, 2017.

[23] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised Representation Learning by Predicting Image Rotations. In *ICLR 2018*, Vancouver, Canada, April 2018.

[24] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, 2018.

[25] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pages 15535–15545, 2019.

[26] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1920–1929, 2019.

[27] Jong-Chyi Su, Subhransu Maji, and Bharath Hariharan. When does self-supervision improve few-shot learning? *arXiv preprint arXiv:1910.03560*, 2019.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[29] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks, 2017.

[30] Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5d heatmap regression, 2018.